

16. Single-Factor ANOVA

SLR: one quantitative predictor, X

MLR: several quantitative predictors, X_1, \dots, X_{p-1}

One-way ANOVA: concerns regression on a *factor* – a predictor with a finite discrete set of *levels*

Example data set: Annual productivity improvements for a sample of firms by level of expenditure on research and development.

i	j					
	1	2	3	4	5	6
Low	7.6	8.2	6.8	5.8		
Moderate	6.7	8.1	9.4	8.6	7.8	7.7
High	9.7	8.5	10.1			

Sampling Plans

- (a) *Experimental factor:* Designed or controlled experiment – random sample at each factor level (stratified random sampling)
- (b) *Classification factor:* Observational study – random sample from the population, classified by factor levels

Notation

- Y_{ij} = j th response at i th factor level
 $i = 1, \dots, r$ (r factor levels)
 $j = 1, \dots, n_i$ (possibly unequal sample sizes)
- $n_T = \sum n_i$ = total sample size
- $Y_{i.} = \sum_j Y_{ij}$ = i th sample total
 $\bar{Y}_{i.} = i$ th sample mean
 $s_i^2 = i$ th sample variance

Cell Means Model

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon \sim N(0, \sigma^2)$$

- $E(Y_{ij}) = \mu_i$
i.e. the mean depends on the factor level.
- Variability about the mean described by a normal distribution, $N(0, \sigma^2)$ – same for all factor levels.
- Same as for SLR and MLR!

Matrix Formulation

e.g. $r = 3, n_1 = 2, n_2 = 3, n_3 = 2$

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{31} \\ Y_{32} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{31} \\ \epsilon_{32} \end{bmatrix},$$

Least Squares Estimation

- Minimize

$$Q = \sum_{i=1}^r \sum_{j=1}^{n_i} (Y_{ij} - E(Y_{ij}))^2$$

with respect to the parameters $\beta = (\mu_1, \dots, \mu_r)'$.

Notice that

$$\begin{aligned} Q &= \sum_{i=1}^r \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2 \\ &= \sum_{j=1}^{n_1} (Y_{1j} - \mu_1)^2 + \sum_{j=1}^{n_2} (Y_{2j} - \mu_2)^2 \\ &\quad + \dots + \sum_{j=1}^{n_r} (Y_{rj} - \mu_r)^2. \end{aligned}$$

It follows that

$$\frac{\partial Q}{\partial \mu_i} = -2 \sum_{j=1}^{n_i} (Y_{ij} - \mu_i).$$

Normal Equations

$$\sum_{j=1}^{n_i} (Y_{ij} - \hat{\mu}_i) = 0 \Rightarrow \hat{\mu}_i = \bar{Y}_i.$$

Fitted Values

$$\hat{Y}_{ij} = E(\widehat{Y}_{ij}) = \hat{\mu}_i = \bar{Y}_i.$$

Residuals

$$e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_i.$$

Matrix Formulation of Least Squares

The matrix formula still holds!

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n_1 & 0 & 0 \\ 0 & n_2 & \dots \\ 0 & \dots & n_r \end{bmatrix} = \text{diag}(n_i)$$

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} Y_{1\cdot} \\ Y_{2\cdot} \\ \vdots \\ Y_{r\cdot} \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \bar{Y}_{1\cdot} \\ \bar{Y}_{2\cdot} \\ \vdots \\ \bar{Y}_{r\cdot} \end{bmatrix} = \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \vdots \\ \hat{\mu}_r \end{bmatrix}$$

7

ANOVA

- Overall total and mean

$$Y_{..} = \sum_{i=1}^r \sum_{j=1}^{n_i} Y_{ij} \quad \bar{Y}_{..} = \frac{1}{n_T} \sum_{i=1}^r \sum_{j=1}^{n_i} Y_{ij}$$

- Decomposition of total deviation

total deviation = deviation of group i + residual

$$Y_{ij} - \bar{Y}_{..} = (\bar{Y}_{i\cdot} - \bar{Y}_{..}) + (Y_{ij} - \bar{Y}_{i\cdot})$$

- Decomposition of total sum of squares

$$\begin{aligned} \text{SSTO} &= \text{SSTR} + \text{SSE} \\ \sum_{i=1}^r \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 &= \sum_{i=1}^r n_i (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2 \\ &\quad + \sum_{i=1}^r \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot})^2 \end{aligned}$$

8

- The *treatment sum of squares*, $SSTR$, corresponds SSR in the regression setting.

- The SSE can be calculated from the treatment group sample variances

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i.)^2$$

Hence

$$SSE = \sum_{i=1}^r (n_i - 1) s_i^2$$

The MSE is unbiased

- The MSE is a weighted average of the sample variances

$$\begin{aligned} MSE &= \frac{\sum_{i=1}^r (n_i - 1) s_i^2}{n_T - r} = \sum_{i=1}^r \frac{n_i - 1}{n_T - r} s_i^2 \\ &= \text{pooled variance estimate} \end{aligned}$$

- Since each sample variance is unbiased for σ^2 , it follows that the MSE is unbiased.

$$\begin{aligned} E(MSE) &= \sum_{i=1}^r \frac{n_i - 1}{n_T - r} E(s_i^2) \\ &= \sum_{i=1}^r \frac{n_i - 1}{n_T - r} \sigma^2 \\ &= \sigma^2 \end{aligned}$$

One-Way ANOVA Table:

Source variat.	Sum of Squares (SS)	df	mean SS
Trts.	$SSTR = \sum_i n_i (\bar{Y}_i. - \bar{Y}..)^2$	$r - 1$	$\frac{SSTR}{r - 1}$
Error	$SSE = \sum_{ij} (Y_{ij} - \bar{Y}_i.)^2$	$n_T - r$	$\frac{SSE}{n - r}$
Total	$SSTO = \sum_{ij} (Y_{ij} - \bar{Y}..)^2$	$n_T - 1$	

MSTR is unbiased if H_0 is true!

- Suppose that $H_0 : \mu_1 = \dots = \mu_r$ is true. Let μ denote the common population mean.
- Suppose also that the sample sizes are equal, $n_1 = \dots = n_r = n$, say.
- Then, the sample means, $\bar{Y}_1, \dots, \bar{Y}_r$, are a sample from $N(\mu, \sigma^2/n)$. (Why?)
- The sample variance of the sample means is unbiased for σ^2/n . Hence

$$E(\text{MSTR}) = \frac{n}{r-1} \sum_{i=1}^r (\bar{Y}_i - \bar{Y}_{..})^2 = n \times \frac{\sigma^2}{n} = \sigma^2$$

- If H_0 is false

$$E(\text{MSTR}) = \sigma^2 + \frac{n \sum (\mu_i - \mu)^2}{r-1} > \sigma^2$$

Application to Productivity Data

Data summary

Group	n_i	$\bar{Y}_{i.}$	s_i
Low	4	7.10	1.04
Moderate	6	8.05	0.91
High	3	9.43	0.83

- Overall mean

$$\bar{Y}_{..} = \frac{4(7.10) + 6(8.05) + 3(9.43)}{4 + 6 + 3} = \frac{105}{13} = 8.08$$

- Treatment sum of squares

$$\begin{aligned} \text{SSTR} &= 4(7.10 - 8.08)^2 + 6(8.05 - 8.08)^2 \\ &\quad + 3(9.43 - 8.08)^2 \\ &= 9.34 \end{aligned}$$

- Error sum of squares

$$\text{SSE} = 3(1.04)^2 + 5(0.91)^2 + 2(0.83)^2 = 8.76$$

ANOVA table

Source of variation	Sum of Squares (SS)	df	mean SS
Groups	SSTR = 9.34	2	4.671
Error	SSE = 8.76	10	0.876
Total	SSTO = 18.10	12	

- F-test for no difference between the groups (or no group effect)

$$F^* = \frac{MSTR}{MSE} = \frac{4.671}{0.876} = 5.33$$

- P-value

$$P\{F(2, 10) \geq 5.33\} = 0.0266$$

One-Way ANOVA in R

```

> y <- c(7.6,8.2,6.8,5.8,...,7.7,9.7,8.5,10.1)
> group <- c("L","L","L",..., "M","M","H","H","H")
> prod.df <- data.frame(y,group)
> grpdata <- split(prod.df$y,prod.df$group)
> grpdata
$H
[1] 9.7 8.5 10.1
$L
[1] 7.6 8.2 6.8 5.8
$M
[1] 6.7 8.1 9.4 8.6 7.8 7.7
> sapply(grpdata,mean)
      H      L      M
9.433333 7.100000 8.050000
> sapply(grpdata,sd)
      H      L      M
0.8326664 1.0392305 0.9093954
> anova(lm(y~group,data=prod.df))
Analysis of Variance Table

      Df Sum Sq Mean Sq F value Pr(>F)
group      2  9.3414  4.6707  5.3308 0.02656 *
Residuals 10  8.7617  0.8762

```

Factor Effects Model

$$\begin{aligned}\mu_i &= \mu. + (\mu_i - \mu.) \\ &= \mu. + \tau_i \quad i = 1, \dots, r\end{aligned}$$

- τ_i is called the *i*th treatment effect
- There is one redundant parameter!
- Homogeneity hypothesis \iff No treatment effects

$$\mu_1 = \mu_2 = \dots = \mu_r \iff \tau_1 = \tau_2 = \dots = \tau_r = 0$$
- Standard default is to include an “intercept” term in a linear model

15

Identifiability Constraints

- (a) Cell means model: $\mu. = 0 \implies \mu_i = \tau_i$
- (b) Treatment contrasts: Treatment effects are contrasts with level 1 (default in R)
- $$\mu. = \mu_1 \implies \tau_i = \mu_i - \mu_1$$
- (c) Treatment contrasts: Treatment effects are contrasts with level r (default in SAS)
- $$\mu. = \mu_r \implies \tau_i = \mu_i - \mu_r$$
- (d) Sum constraint: Treatment effects measure deviation from center

$$\mu. = \frac{1}{r} \sum_{i=1}^r \mu_i \implies \sum_{i=1}^r \tau_i = 0$$

16

Matrix Formulation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

e.g. $r = 3, n_1 = 2, n_2 = 3, n_3 = 2$

Treatment Contrasts Model

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu. \\ \tau_2 \\ \tau_3 \end{bmatrix} = \begin{bmatrix} \mu. \\ \mu. \\ \mu. + \tau_2 \\ \mu. + \tau_2 \\ \mu. + \tau_2 \\ \mu. + \tau_3 \\ \mu. + \tau_3 \end{bmatrix}$$

Sum Constraints Model

$$\mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \mu. \\ \tau_1 \\ \tau_2 \end{bmatrix} = \begin{bmatrix} \mu. + \tau_1 \\ \mu. + \tau_1 \\ \mu. + \tau_2 \\ \mu. + \tau_2 \\ \mu. + \tau_2 \\ \mu. - \tau_1 - \tau_2 \\ \mu. - \tau_1 - \tau_2 \end{bmatrix}$$

17

Productivity Data

```
> grpdata <- split(prod.df$y, prod.df$group)
> sapply(grpdata, mean)
      H      L      M
9.433333 7.100000 8.050000
```

Note that factor levels are ordered alphabetically

$$1 = H, \quad 2 = L, \quad 3 = M$$

Treatment Contrasts Model

```
> model.l.1 <- lm(y~group, data=prod.df)
> summary(model.l.1)
```

Coefficients:

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.43333    0.5404    17.456 8.08e-09 ***
groupL      -2.33333    0.7149    -3.264 0.00852 **
groupM      -1.38333    0.6619    -2.090 0.06313 .
```

$$\hat{\mathbf{Y}} = 9.43 - 2.33 I\{L\} - 1.38 I\{M\}$$

18

Sum Constraints Model

```

> options(contrasts=c("contr.sum", "contr.poly"))
> model.2 <- lm(y~group, data=prod.df)
> summary(model.2)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.1944    0.2702  30.326 3.56e-11 ***
group1         1.2389    0.4128   3.002  0.0133 *
group2        -1.0944    0.3821  -2.864  0.0168 *

```

$$\hat{Y} = 8.19 + 1.24 I\{H\} - 1.09 I\{L\} - 0.14 I\{M\}$$

$$\bar{Y}_1 = 8.19 + 1.24 = 9.43$$

$$\bar{Y}_2 = 8.19 - 1.09 = 7.10$$

$$\bar{Y}_3 = 8.19 - 1.24 + 1.09 = 8.05$$

Pairwise Comparisons

- Sample means are independent normal variables

$$\bar{Y}_i - \bar{Y}_j \sim N \left[\mu_i - \mu_j, \sigma^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right) \right]$$

$$\frac{(\bar{Y}_i - \bar{Y}_j) - (\mu_i - \mu_j)}{\sqrt{\sigma^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \sim N(0, 1)$$

$$\frac{(\bar{Y}_i - \bar{Y}_j) - (\mu_i - \mu_j)}{\sqrt{s^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \sim t(n_T - r)$$

where $s^2 = \text{MSE}$.

- Test $H_0 : \mu_i = \mu_j$ using

$$t^* = \frac{\bar{Y}_i - \bar{Y}_j}{\sqrt{s^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}}$$

Productivity Data

Comparison	Difference	Std.Error	t^*	P-value
H-M	1.38	0.662	2.08	0.0617
H-L	2.33	0.715	3.26	0.0076
M-L	0.95	0.604	1.57	0.1447

Comments

- Same as two-sample t-tests, except pooled variance from all samples.
- If the sample sizes are equal then there is a common *least significant difference* (LSD)
- Multiple comparisons!